This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

Mask to Concept: Auto-Promptable SAM3 via Efficient Test-Time Concept Embedding Search for Few-Shot Annotation

Quan Zhou[†1], Shaoqing Zhai[†1], Qiang Hu[(✉)2], Jia Chen[3], Qiang Li[2], and Zhiwei Wang[(✉)2]

[1] Wuhan University of Technology
[2] Huazhong University of Science and Technology
[3] Changzhou United lmaging Healthcare Surgical Technology Co. Ltd.
{zhouquan910, zhaishaoqing}@whut.edu.cn, {huqiang77, zwwang}@hust.edu.cn

Abstract. Transforming foundation segmentation models from human-prompted tools into auto-promptable annotators is critical for scalable medical data annotation. Current methods commonly depend on external feature matchers or auxiliary networks to automate geometric prompting, but introducing architectural overhead and limiting performance scalability. Although SAM3 natively supports concept segmentation via reusable text prompts, its direct use in medical imaging is hindered by a lack of fine-grained clinical knowledge and the ambiguity of human-written descriptions. In this work, we propose Mask to Concept (M2C), an efficient framework that adapts SAM3 for medical few-shot annotation *without* external modules, parameter retraining, or manual text engineering. Using only a few labeled images, M2C enables SAM3 to automatically search for transferable visual concepts entirely within its frozen architecture: it initializes a learnable concept embedding, uses it to prompt segmentation, and updates the embedding by gradients of minimizing the concept segmentation error. We further introduce a Hybrid Uncertainty Estimation (HUE) module that calculates the prediction entropy and maps concept predictions back to the box prompts, measuring concept-geometry prompting inconsistency. Highly uncertain samples are flagged actively for human correction, and the corrected masks are then fed back to M2C to continuously search for more precise concept embeddings, forming a self-enhancing annotation loop with minimal expert effort. Experiments on medical segmentation benchmarks show that our method achieves SOTA few-shot segmentation performance and outstanding annotation efficiency, offering a practical and efficient pathway toward scalable medical image labeling. Codes are at https://github.com/Huster-Hq/M2C.

Keywords: Segment Anything Model · Few-shot Segmentation · Automatic Annotation System.

[†] Equal contribution; ✉ Corresponding author.

1 Introduction

High-quality, large-scale annotations are essential for modern medical AI, yet obtaining pixel-level segmentation labels is prohibitively expensive [1,2]. This annotation bottleneck has motivated research into training-free tools. Models like SAM [3,4] and its clinical variants (*e.g.*, MedSAM [5]) offer a promising start, enabling segmentation from simple geometric prompts. However, they operate as per-image interactive assistants, requiring manual input for each case and limiting scalability for large-scale annotation.

To amortize human effort, recent works seek to develop automatic annotators driven by a few labeled samples. These works can be categorized into two types: The first category is *training-based methods* [6,7,8,9,10], which heavily rely on pre-training under the few-shot segmentation (FSS) paradigm on large-scale medical datasets. However, their performance tends to be constrained within the training data distribution, exhibiting poor generalization and flexibility. Conversely, *training-free methods* typically leverage SAM by introducing automated geometric prompt generation mechanism to drive the annotation process. However, they rely on external modules, such as separate feature matching modules [11,12,13,14] or auxiliary networks [15,16,17], to establish cross-instance correspondence. This not only complicates the pipeline but often fails to fully leverage SAM's internal reasoning, resulting in inconsistent quality: performance lags behind per-image manual prompting and rarely improves with more references.

The release of SAM3 [18], with its native support for promptable concept segmentation using texts, marks a pivotal shift. By moving from case-specific geometric prompts to reusable semantic concepts carried by text prompts, it opens a direct pathway toward a concept-driven annotation assistant. This raises our core research question: *Can we unlock SAM3's concept segmentation to build a cross-instance few-shot annotator entirely within its native capabilities?* However, achieving this in medical imaging is challenging. SAM3 lacks inherent knowledge of fine-grained clinical concepts (*e.g.*, specific lesion subtypes), and human-written text prompts are often too vague for reliable [19,20], automated labeling across a diverse dataset.

To address these challenges, we make the following major contributions: **(1)** We introduce Mask to Concept (M2C), a novel mechanism that enables SAM3 to automatically search for medical visual concepts from annotated masks, bypassing ambiguous text prompts and adapting to the medical domain without FSS-Pretraining and SAM3-retraning. Instead of manual prompt engineering, M2C initializes an efficient *learnable* concept embedding and refines it through a differentiable pathway: given a few annotated images, the embedding prompts SAM3 to predict masks, and the prediction error is backpropagated to update the embedding. This process is fast, memory-efficient, and grows more accurate with more examples, naturally supporting incremental few-shot learning. **(2)** We further integrate M2C into a human-in-the-loop annotation system, where the M2C embedding enables SAM3 to annotate the rest unlabeled data autonomously. To identify samples requiring human verification, we design a hybrid uncertainty estimation (HUE) module to evaluate prediction uncertainty. Low-uncertainty

samples are auto-labeled, while high-uncertainty ones are sent for human verification. Newly corrected masks then refine the concept embedding, closing an active learning loop that continuously improves annotation quality with minimal human effort. **(3)** Extensive experiments demonstrate that M2C achieves SOTA FSS performance: under the one-shot setting, M2C outperforms existing methods by at least 4.2% and 11.8% in Dice on Kvasir-SEG and ISIC-2017, respectively. Moreover, the proposed annotation system exhibits superior efficiency and data generalization on both datasets, showing its practical significance.

2 Method

2.1 Overview

As illustrated in Fig. 1, our human-in-the-loop annotation framework based on SAM3 is empowered by two iterative modules: Mask to Concept (M2C) and Hybrid Uncertainty Estimation (HUE). Given a specific medical dataset, users first annotate the masks for one or several samples using SAM3 with simple geometric prompts (*e.g.*, boxes or points). Then M2C rapidly learns a ***concept embedding*** specific to the dataset derived from these mask annotations, allowing SAM3 to automatically segment the remaining unlabeled samples via the concept prompting mechanism. Next, HUE evaluates these predicted masks and ranks samples based on the estimated uncertainty scores. High-uncertainty samples will be proactively selected for manual label correction, with the corrected labels fed back to M2C for continuous search of the concept embedding. Low-uncertainty samples can be optionally verified and removed from the unlabeled sample pool. These two modules are iterated until the entire dataset is annotated, enabling scalable accurate annotation with cost-effective labor costs.

2.2 Mask-to-Concept for Concept Searching

To overcome SAM3's inherent reliance on text descriptions for concept segmentation and enable its flexible adaptation to arbitrary scenarios, we propose Mask-to-Concept (M2C). M2C learns a concept embedding from a few labeled samples, transforming SAM3 into a dataset-specific automatic segmenter while keeping its backbone parameters strictly frozen.

Specifically, given a medical image x and a human-written text prompt T (*e.g.*, a category name or a sentence), SAM3 originally utilizes a text encoder to obtain a concept embedding $\texttt{TEnc}(T)$, and then predicts a segmentation mask $y^{\text{pred}} = \texttt{SAM}(x, \texttt{TEnc}(T))$. To search a better concept, we introduce a learnable vector $\mathbf{E}$ to revise the original concept embedding, and thus the segmentation process can be re-formulated as $y^{\text{pred}} = \texttt{SAM}(x, \texttt{TEnc}(T) + \mathbf{E})$.

With this modification, a continual searching strategy is proposed to iteratively refine $\mathbf{E}$, enabling it to progressively adapt to remaining hard samples. At the t-th iteration, given a set of human-annotated samples $\mathcal{D}_{t-1}^{\text{label}} =$

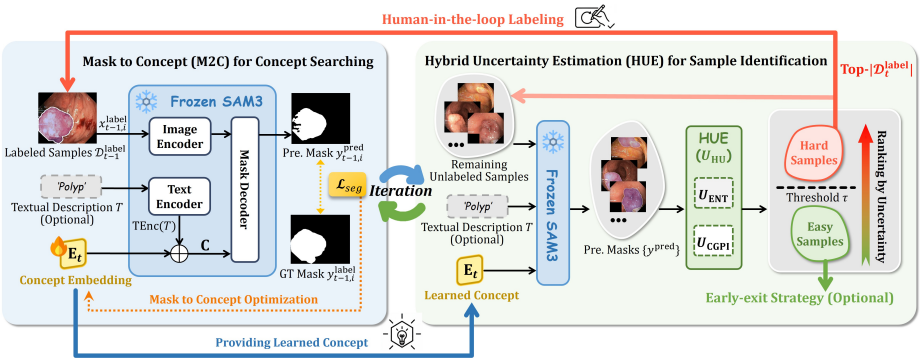


Fig. 1. Overview of the proposed human-in-the-loop annotation framework. It consists of two core iterative modules: Mask to Concept (M2C) and Hybrid Uncertainty Estimation (HUE). During each iteration, given a few human-labeled samples from the previous iteration, M2C employs a continuous searching on the dataset-specific concept embedding to automatically segment the remaining unlabeled samples. Following this, HUE estimates the uncertainties of these predictions and applies tailored strategies to the samples based on the uncertainty as they transition into the next iteration.

$\{(x^{\text{label}}_{t-1,i}, y^{\text{label}}_{t-1,i})\}_{i=1}^{|\mathcal{D}^{\text{label}}_{t-1}|}$ from the previous iteration, we freeze the entire SAM3 model and optimize only the learnable embedding. Here, $x^{\text{label}}_{t-1,i}$ and $y^{\text{label}}_{t-1,i}$ represent the i-th image and labeled mask, respectively. Initialized with $\mathbf{E}_{t-1}$ from the prior iteration, the embedding is optimized end-to-end by minimizing the segmentation loss via gradient descent:

$$\mathbf{E}_t = \mathbf{E}_{t-1} - \eta \nabla_{\mathbf{E}_{t-1}} \left(\frac{1}{|\mathcal{D}^{\text{label}}_{t-1}|} \sum_{i=1}^{|\mathcal{D}^{\text{label}}_{t-1}|} \mathcal{L}_{\text{seg}} \left(y^{\text{pred}}_{t-1,i}, y^{\text{label}}_{t-1,i} \right) \right), \tag{1}$$

where $y^{\text{pred}}_{t-1,i} = \texttt{SAM}(x^{\text{label}}_{t-1,i}, \texttt{TEnc}(\texttt{T}) + \mathbf{E}_{\texttt{t}-1})$, and η is the searching step size. The objective function $\mathcal{L}_{\text{seg}}$ is the native training loss of SAM3, comprising the Intersection over Union (IoU) loss and cross-entropy loss. Note that, with this continual searching strategy, the text prompt T is no longer necessary.

2.3 Hybrid Uncertainty Estimation for Sample Identification

To efficiently distinguish hard samples requiring subsequent human intervention from easy samples with already satisfactory annotation quality, we introduce an automatically calculated hybrid uncertainty metric, denoted as $U_{\text{HU}}(\cdot)$. The metric quantifies the model's predicted map y^{pred} (we omit indexes of iteration and sample subscripts for brevity) by integrating two complementary sub-metrics:

(1) **Prediction Entropy**. We calculate the pixel-wise entropy of the predicted map to capture decision boundary ambiguity. Let $p_i \in [0, 1]$ denote the value of the i-th pixel in y^{pred}, its Shannon entropy is defined as: $H(p_i) =$

$-\big(p_i \log p_i + (1-p_i)\log(1-p_i)\big)$. To strictly focus on informative and uncertain regions, we define a set of ambiguous pixels based on their local entropy thresholds: $\Omega = \{i \mid H(p_i) > 0.8\}$. Consequently, the final prediction entropy for y^{pred} is formulated as the mean entropy over this set: $U_{\text{ENT}}(y^{\text{pred}}) = \frac{1}{|\Omega|}\sum_{i\in\Omega} H(p_i)$.

(2) **Concept-Geometry Prompting Inconsistency**. We transform the initial concept-driven predicted mask y^{pred} into a bounding box. The box serves as a geometric prompt for SAM3 to generate a geometry-driven mask y^{box}. The uncertainty derived from this cross-prompt mechanism is defined by the IoU score between the two masks: $U_{\text{CGPI}}(y^{\text{pred}}) = 1 - \texttt{IoU}(y^{\text{pred}}, y^{\text{box}})$.

The final hybrid uncertainty score for a given predicted map y^{pred} is formulated as the sum of these two sub-metrics: $U_{\text{HU}}(y^{\text{pred}}) = U_{\text{ENT}}(y^{\text{pred}}) + U_{\text{CGPI}}(y^{\text{pred}})$.

After obtaining the hybrid uncertainties, we rank the predictions based on uncertainties and threshold them into hard and easy samples via a threshold τ. The top-$|\mathcal{D}_t^{\text{label}}|$ hard samples are routed for manual label correction, serving as the labeled samples $\mathcal{D}_t^{\text{label}}$ to support the continuous search for the concept embedding in the $(t+1)$-th iteration. While the remaining hard samples are treated as the remaining unlabeled data and are fed into the next iteration. Meanwhile, an optional early-exit strategy applied to the easy samples, completing their annotation process early and excluding them from subsequent iterations.

3 Experiments

3.1 Experimental Setup

Baselines. We compare our M2C with six state-of-the-art (SOTA) few-shot segmentation (FSS) methods, categorized into training-based methods (Multiverseg [8], Tyche [7]) and training-free[1] methods (SPFS-SAM [15], MAUP [13], Matcher [11], ProtoSAM [12]) based on their requirement for FSS pre-training.

Datasets. We conduct experiments on two medical datasets: Kvasir-SEG [21] for endoscopic polyp segmentation and ISIC-2017 [22] for skin lesion segmentation. These two datasets are both unseen during the pre-training of SAM3, which can effectively verify the model's zero-shot generalization capabilities. For our evaluation protocol, we utilize the entire Kvasir-SEG dataset, comprising 1,000 images. Conversely, for the ISIC-2017 dataset, we strictly confine our evaluation to the official test set of 600 images. This restriction is implemented to prevent data leakage and to ensure a fair comparison with Tyche and Multiverseg, which were exposed to the ISIC training split during its pre-training.

Evaluation Metrics. We employ Dice Similarity Coefficient to quantitatively evaluate segmentation performance.

[1] Here, 'training-free' specifically means that the parameters of SAM3 remain unchanged and the approach does not require training by the FSS paradigm.

Table 1. Quantitative comparison with SOTA FSS methods on Kvasir-SEG and ISIC-2017 datasets. We report the mean and standard deviation ($\mu \pm \sigma$) for the Dice metric as percentages(%). Methods are divided into training-based (*top*) and training-free (*bottom*). The best results are highlighted in **bold**, and the second-best are underlined.

Methods	Kvasir-SEG			ISIC-2017		
	1-shot	5-shot	10-shot	1-shot	5-shot	10-shot
Tyche [7]	26.9±8.2	43.1±2.0	49.9±1.5	37.2±15.9	68.1±3.5	69.7±2.8
Multiverseg [8]	35.8±6.7	48.5±4.6	50.5±3.1	46.3±5.1	70.8±3.5	71.6±2.3
ProtoSAM [12]	72.1±4.8	–	–	64.1±5.3	–	–
Matcher [11]	67.1±13.5	68.3±1.9	69.2±4.9	54.2±11.3	58.8±8.2	58.3±2.3
SPFS-SAM [15]	70.4±7.4	78.5±1.1	80.4±0.6	52.3±8.4	72.3±2.5	74.9±1.2
MAUP [13]	46.9±5.7	60.6±3.9	58.3±6.3	53.0±8.5	55.7±6.6	59.3±4.4
M2C (Ours)	**76.3±8.0**	**80.2±4.0**	**82.8±4.4**	**75.9±8.4**	**79.2±2.5**	**82.1±0.5**

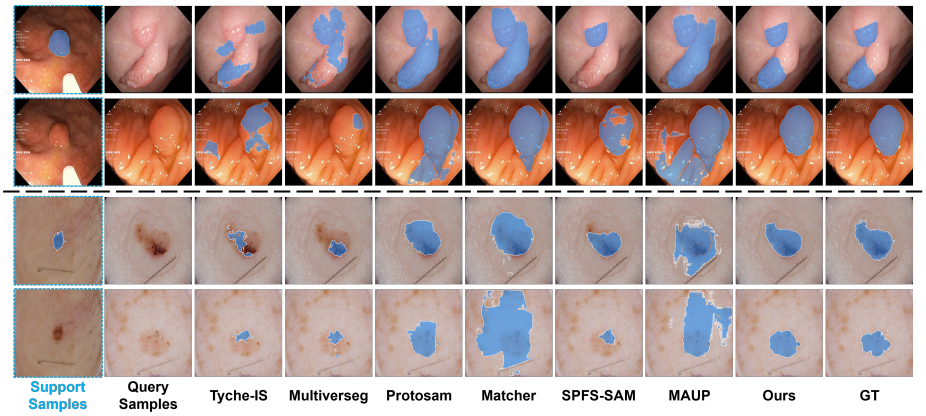


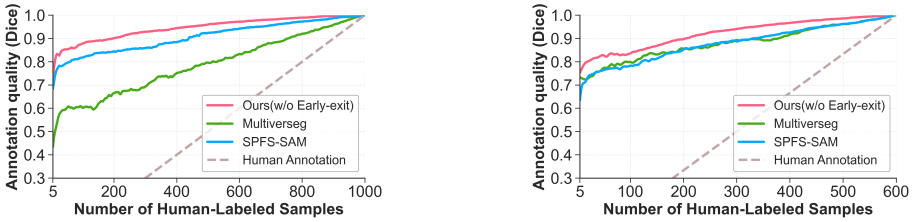
Fig. 2. Visualization of 1-shot results on Kvasir-SEG (*top*) and ISIC-2017 (*bottom*).

3.2 Implementation Details

We implement our method based on the official pre-trained weight of SAM3 [18]. To ensure a fair comparison, all SAM-based baselines (*i.e.*, ProtoSAM, Matcher, SPFS-SAM, and MAUP) utilize SAM-H [3]. During the mask to concept continual searching, we freeze the entire SAM3 backbone and exclusively optimize the concept embedding. The embedding optimization is executed on a single NVIDIA RTX 4090 GPU, optimized by AdamW [23] optimizer. The searching step size η is set to 5×10^{-4} and follows a cosine decay strategy down to 5×10^{-6}. We set the uncertainty threshold τ to 0.6 to distinguish between simple and hard samples. The concept embedding is initialized as an all-zero vector.

3.3 Comparison in Few-shot Segmentation

We compare our M2C with six FSS SOTAs on both the Kvasir-SEG and ISIC-2017 datasets independently, and all these methods are evaluated directly with-



(a) Kvasir-SEG (b) ISIC-2017

Fig. 3. Annotation efficiency curves on Kvasir-SEG and ISIC-2017 datasets.

out further fine-tuning. In our FSS evaluation protocol, the dataset is split into support and query sets using a $1:9$ ratio. For K-shot segmentation, we test performance on the query set by forming episodes where one query sample is paired with K randomly selected support samples. To ensure statistical reliability, each query image undergoes five independent tests with different support sets, and the mean performance and standard deviation of these five rounds are reported.

Quantitative Results. As shown in Table 1, the training-based methods, Tyche and Multiverseg, typically exhibit limited performance. Despite undergoing FSS-paradigm pre-training on large-scale medical datasets, they still demonstrate poor generalization capabilities. Conversely, our M2C significantly outperforms all competing methods across both datasets and under all shot settings. Particularly in the 1-shot setting, M2C outperforms the second-best method, ProtoSAM, by a Dice improvement of 4.2% (76.3% *vs.* 72.1%) on Kvasir-SEG and 11.8% (75.9% *vs.* 64.1%) on ISIC-2017, respectively. This superiority is attributed to our M2C mechanism, which can learn precise concept representations from extremely limited samples during test time, thereby facilitating robust and efficient adaptation to diverse scenarios.

Qualitative Results. Fig. 2 visualizes the qualitative comparisons of four cases from the Kvasir-SEG and ISIC 2017 under the 1-shot setting. As illustrated, our M2C can precisely segment ambiguous and complex boundaries of query images, even under significant visual variations. In contrast, other training-free methods relying on SAM's geometric prompting mechanism are limited by the SAM's native capabilities, and the strategy of generating geometric prompts through feature matching struggles to overcome challenges such as blurry boundaries, background interference, and large cross-image visual variations.

3.4 Comparison in Annotation Efficiency

Motivated by the potential of these methods in automated data annotation, we compare the annotation efficiency of our M2C-based system with Multiverseg and SPFS-SAM. We quantify this efficiency by measuring the overall dataset annotation quality under different numbers of human-labeled samples provided. Specifically, the overall annotations include the human-labeled samples simulated using GT masks, and the remaining samples predicted by the model. Notably, to ensure a fair comparison with methods lacking automatic annotation quality estimation mechanisms, we disable the early-exit strategy for easy samples.

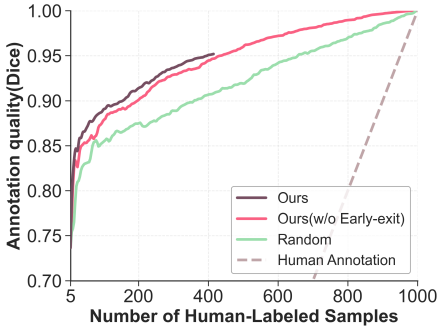


Fig. 4. Ablation of HUE on Kvasir-SEG.

Table 2. Ablation of two uncertainty sub-metrics (U_{ENT}, U_{CGPI}) on Kvasir-SEG. 'Num' and 'Ratio' represent the number and proportion of manually annotated samples required to achieve a Dice score of 90%, respectively, out of 1,000 images.

Sub-metrics		**Ann. Cost**	
U_{ENT}	U_{CGPI}	**Num**	**Ratio**
✓		345	34.5%
	✓	230	23.0%
✓	✓	**155**	**15.5%**

In Fig. 3, we plot the overall annotation quality curves as the number of annotated samples increases by intervals of 5 (*i.e.*, $|\mathcal{D}_t^{\text{label}}| = 5$). As observed, our method demonstrates the best efficiency, achieving satisfactory annotation quality with limited human annotation budget. Furthermore, our method maintains a consistent superiority on different scenarios, showing the generalizability as an automated annotation tool and underscoring its significant practical value.

3.5 Ablation Study

Effect of HUE on Annotation Efficiency. To validate the efficacy of the proposed HUE, we conduct an ablation study on Kvasir-SEG comparing three configurations: (1) *Random*: We randomly select samples for manual annotation during each iteration. (2) *Ours (w/o Early-exit)*: We use HUE to select hard samples and employ active manual annotation, but do not implement the early-exit strategy for easy samples. (3) *Ours*: Similar to *Ours (w/o Early-exit)* but implement the early-exit strategy for easy samples. We plot the annotation efficiency curves for all configurations in Fig. 4. *Ours (w/o Early-exit)* surpasses the *Random* configuration, verifying that our hybrid uncertainty metric can retrieve valuable samples to maximize dataset quality under fixed human annotation budgets. Additionally, the complete configuration *Ours* achieves the most satisfactory annotation quality with minimal manual effort. This demonstrates the vital role of early-exit for easy samples, ensuring they undergo precise segmentation through concept-aligned embeddings before continuous updates.

Effect of Two Uncertainty Sub-metrics on Annotation Efficiency. Accurate uncertainty estimation is crucial for our human-in-the-loop annotation systems, as it directly dictates the early-exit of easy samples and the routing of hard samples for manual correction. In Table 2, we assess the impact of the two sub-metrics (prediction entropy U_{ENT} and concept-geometry prompting inconsistency U_{CGPI}) on annotation efficiency by measuring the number of manually annotated samples required to achieve a target quality (Dice of 90%). The results show relying solely on U_{ENT} or U_{CGPI} incurs a higher annotation cost than their combined strategy. This demonstrates their complementarity and the importance of integrating them to accurately identify easy and hard samples.

4 Conclusion

In this work, we propose M2C, a novel mechanism designed to efficiently adapt SAM3 to specific medical imaging datasets under few-shot setting. M2C can search the dataset-specific concept embedding from limited annotated masks, marking the first attempt to unlock the native textual concept prompting of SAM3 for medical scenarios without retraining. Furthermore, we integrate M2C with HUE and human interaction to establish a semi-automatic, closed-loop annotation system. Extensive experiments on the Kvasir-SEG and ISIC-2017 datasets demonstrate that M2C significantly outperforms SOTA FSS methods. Moreover, the proposed human-in-the-loop annotation system exhibits superior annotation efficiency and generalization, underscoring its potential as a highly efficient solution for real-world medical data annotation.

Acknowledgments. This work was supported by National Natural Science Foundation of China (No.62401414 and No.62576145), Joint Funds of the Natural Science Foundation of Hubei Province, China (No.2026AFC0648), and research grants from Wuhan United Imaging Healthcare Surgical Technology Co., Ltd.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Nima Tajbakhsh, Laura Jeyaseelan, Qian Li, Jeffrey N Chiang, Zhihao Wu, and Xiaowei Ding. Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation. *Medical image analysis*, 63:101693, 2020.
2. Qiang Hu, Zhenyu Yi, Ying Zhou, Fang Peng, Mei Liu, Qiang Li, and Zhiwei Wang. Sali: Short-term alignment and long-term interaction network for colonoscopy video polyp segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 531–541. Springer, 2024.
3. Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023.
4. Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.
5. Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15(1):654, 2024.
6. Victor Ion Butoi, Jose Javier Gonzalez Ortiz, Tianyu Ma, Mert R Sabuncu, John Guttag, and Adrian V Dalca. Universeg: Universal medical image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21438–21451, 2023.
7. Marianne Rakic, Hallee E Wong, Jose Javier Gonzalez Ortiz, Beth A Cimini, John V Guttag, and Adrian V Dalca. Tyche: Stochastic in-context learning for medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11159–11173, 2024.

8. Hallee E Wong, Jose Javier Gonzalez Ortiz, John Guttag, and Adrian V Dalca. Multiverseg: scalable interactive segmentation of biomedical imaging datasets with in-context guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20966–20980, 2025.
9. Ziming Cheng, Shidong Wang, Tong Xin, Tao Zhou, Haofeng Zhang, and Ling Shao. Few-shot medical image segmentation via generating multiple representative descriptors. *IEEE Transactions on Medical Imaging*, 43(6):2202–2214, 2024.
10. Qiang Hu, Jiajie Wei, Zhenyu Yi, Zhifen Yan, Yingjie Guo, Hongkuan Shi, Ge-Peng Ji, Qiang Li, and Zhiwei Wang. Samix: Reinforcing sam2 with semantic adapter and reference selecting policy for mix-supervised segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17948–17958, 2026.
11. Yang Liu, Muzhi Zhu, Hengtao Li, Hao Chen, Xinlong Wang, and Chunhua Shen. Matcher: Segment anything with one shot using all-purpose feature matching. *arXiv preprint arXiv:2305.13310*, 2023.
12. Lev Ayzenberg, Raja Giryes, and Hayit Greenspan. Protosam: One-shot medical image segmentation with foundational models. *arXiv preprint arXiv:2407.07042*, 2024.
13. Yazhou Zhu and Haofeng Zhang. Maup: Training-free multi-center adaptive uncertainty-aware prompting for cross-domain few-shot medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 326–336. Springer, 2025.
14. Renrui Zhang, Zhengkai Jiang, Ziyu Guo, Shilin Yan, Junting Pan, Xianzheng Ma, Hao Dong, Peng Gao, and Hongsheng Li. Personalize segment anything model with one shot. *arXiv preprint arXiv:2305.03048*, 2023.
15. Qi Wu, Yuyao Zhang, and Marawan Elbatel. Self-prompting large vision models for few-shot medical image segmentation. In *MICCAI workshop on domain adaptation and representation transfer*, pages 156–167. Springer, 2023.
16. Yufei Liu, Haoke Xiao, Jiaxing Chai, Yongcun Zhang, Rong Wang, Zijie Meng, and Zhiming Luo. Synpo: Boosting training-free few-shot medical segmentation via high-quality negative prompts. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 594–603. Springer, 2025.
17. Qiang Hu, Mei Liu, Qiang Li, and Zhiwei Wang. First-frame supervised video polyp segmentation via propagative and semantic dual-teacher network. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025.
18. Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719*, 2025.
19. Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022.
20. Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022.
21. Debesh Jha, Pia H Smedsrud, Michael A Riegler, Pål Halvorsen, Thomas de Lange, Dag Johansen, and Håvard D Johansen. Kvasir-seg: A segmented polyp dataset. In *MultiMedia Modeling: 26th International Conference, MMM 2020, Daejeon, South Korea, January 5–8, 2020, Proceedings, Part II 26*, pages 451–462. Springer, 2020.

22. Noel CF Codella, David Gutman, M Emre Celebi, Brian Helba, Michael A Marchetti, Stephen W Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, et al. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*, pages 168–172. IEEE, 2018.
23. Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.